"Big Data" in NASA and Beyond

L. Harper Pryor - CSC
NAS Division
NASA Ames Research Center
harper.pryor@nasa.gov

2012 Summer Short Course for Earth System Modeling and Supercomputing

“Big Data”

What is it?

Why do you care?

NASA’s Big Data

What's out there?

Where is it?

How do you access/acquire it?

Taming the Big Data Beast

Will it be this…

Or this…

Your Speaker – L. Harper Pryor

CSC's Program Manager for support to NAS

Over 30 years at GSFC supporting major programs in computation and science data processing for earth and space science missions

- Helped transform the NCCS into the NASA Center for Climate Simulation, with an emphasis on data intensive computing
- Coordinated support to major scientific field programs and provided system engineering to high profile projects such as the MAP '05 Hurricane Tracking Project
- Provided System Engineering and technical leadership for computing support to NASA's Global Modeling and Assimilation Office
- Supported many NASA flight and instrument programs, including the Solar Maximum Mission (SMM), Shuttle Solar Backscatter Ultraviolet experiment (SSBUV), Cosmic Background Explorer (COBE), Tropical Rainfall Monitoring Mission (TRMM), Landsat Data Continuity Mission (LDCM), and the Global Precipitation Measuring Mission (GPM)

Acknowledgements

Grateful thanks to the following sources for information used on various of the slides that follow…

- CSC ClimateEdge presentation
- NSF Earthcube presentation
- NASA EOSDIS documentation and websites
- GES DISC training materials
- OSTP press releases and backgrounders

What is the Big Data Problem?

The explosive growth of massive datasets, termed "Big Data", is fueled by pervasive devices observing the natural world, high-fidelity scientific instruments, and mobile and online sensors embedded in our daily lives.

This data deluge exhibits not just volume and velocity but also variability and diversity of structure, completeness and domain.

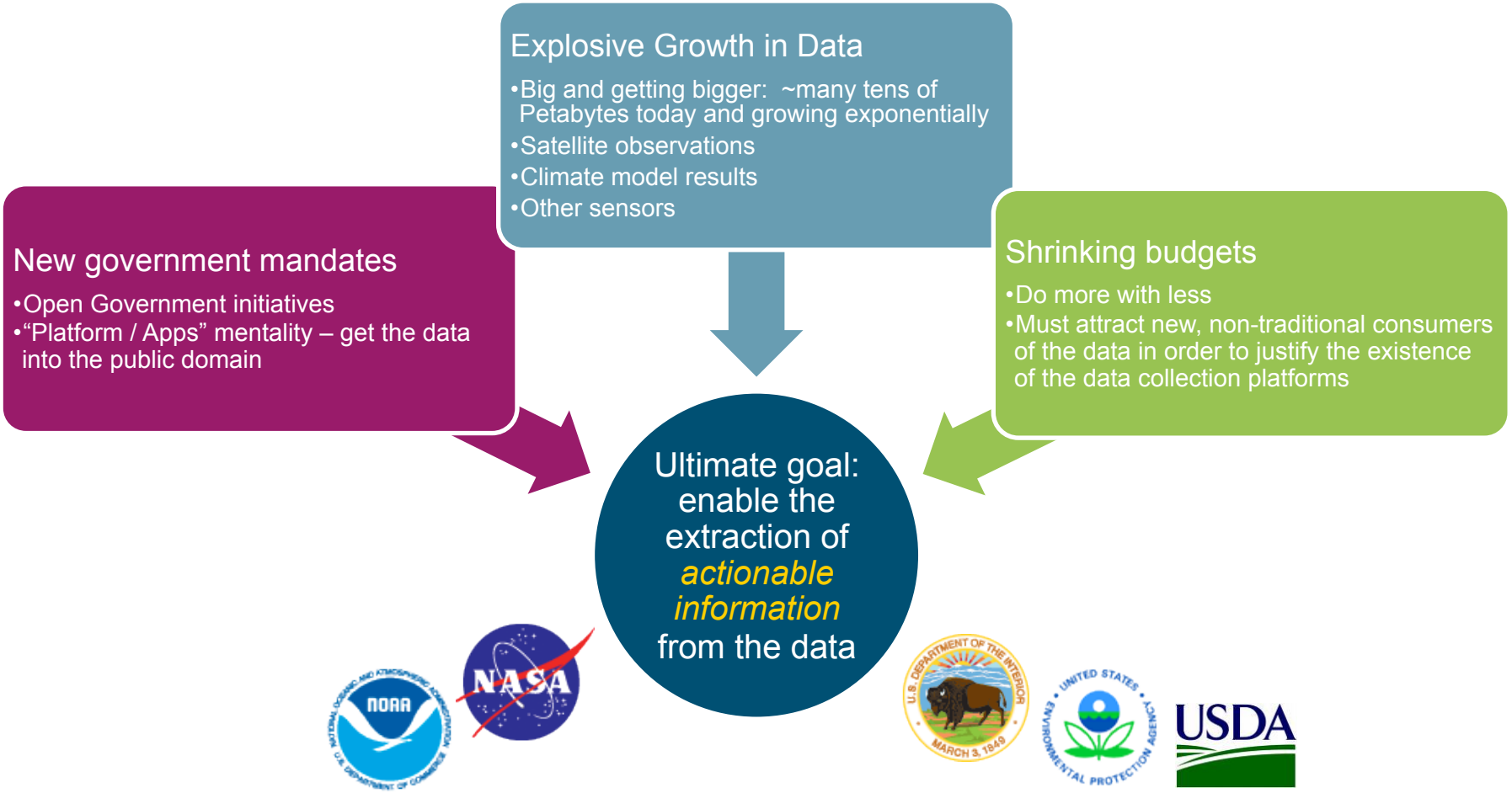
Big Data offers an unprecedented opportunity to understand scientific and social processes, and for deep insight that transcends situational awareness.

However…

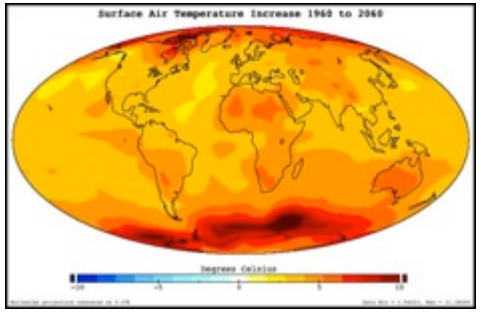
The potential impact of these data is constrained by our ability to rapidly and comprehensively navigate and analyze them

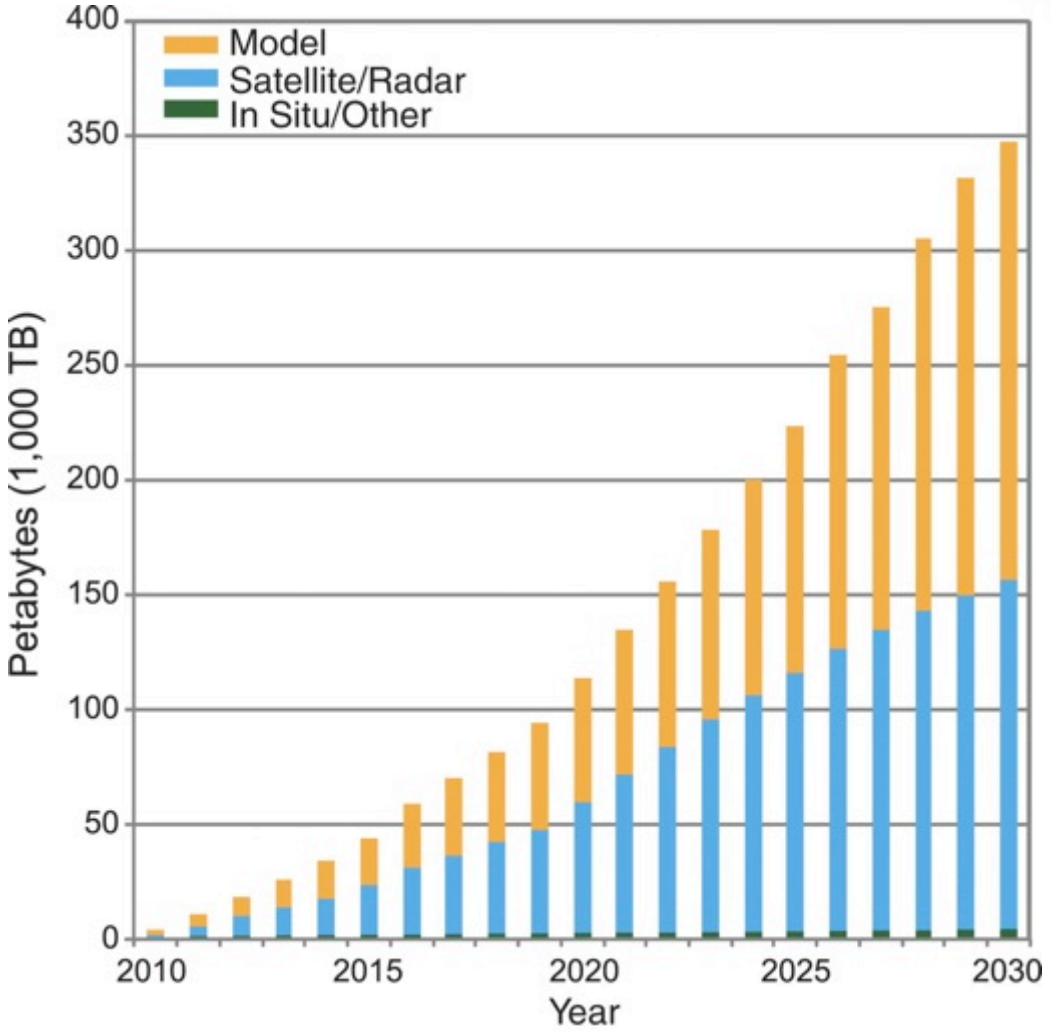
Their size overwhelms traditional data management and analysis techniques, and requires novel algorithms, infrastructure and frameworks to support advanced analytics

Government Agencies Face New Challenges Associated with Climate Change "Big Data"



Climate Change Data Alone is Projected to Grow to Nearly 350 Petabytes by 2030





Big Data is both Size and Speed

Modern geoscience

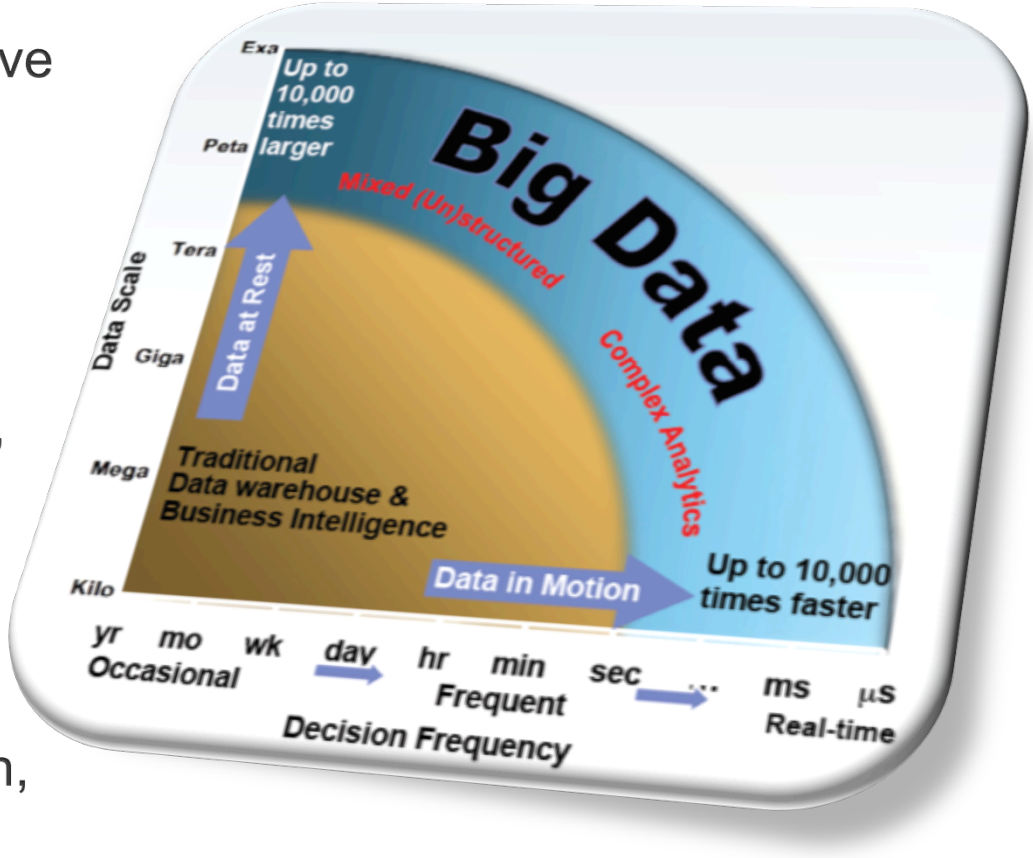
- Data- and compute-intensive
- Integrative, multi-scale

Multi-disciplinary collaborations to address complexity

- Individuals, groups, teams, communities

Sea of Data

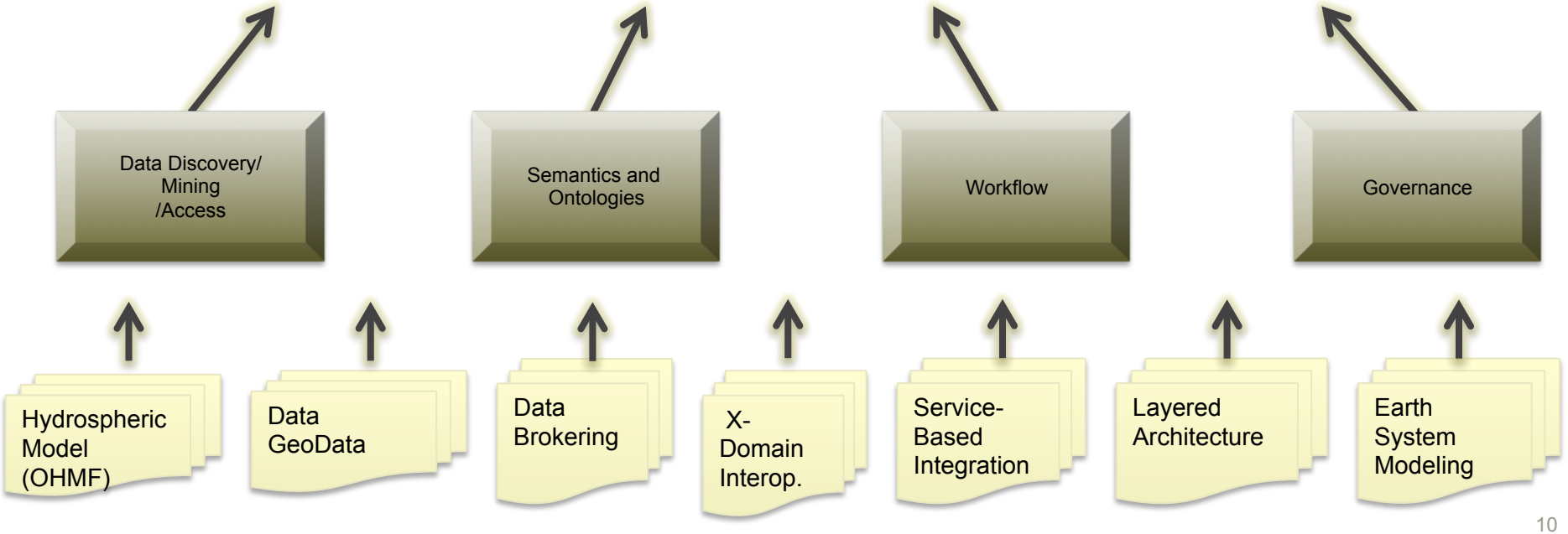
- Age of Observation
- Distributed, central repositories, sensor- driven, diverse, etc.



There is more to Big Data than just data!

EarthCube
National Data Infrastructure for Earth System Science

NSF sponsored effort to collaboratively produce a framework to form an integrated & synergistic path forward to a data and knowledge management system to support 21st Century science



So who is doing something about this?

Everyone!!

Academia – IT Industry – Business - Government

Myriad efforts throughout Government

Programs in just about every department/agency

"Big Data Research and Development Initiative"

-- an overarching administration initiative

Big Data Research and Development Initiative

- Announced by the Obama administration in March 2012
- Six Federal departments/agencies, $200 million in new commitments
- To "greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data"
- Coordinated by OSTP Senior Steering Group on Big Data

"In the same way that past Federal investments in information-technology R&D led to dramatic advances in supercomputing and the creation of the Internet, the initiative we are launching today promises to transform our ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security"

Dr. John P. Holdren, Assistant to the President and Director of the White House Office of Science and Technology Policy

Big Data Research and Development Initiative

NSF and NIH - Joint solicitation for Core Techniques and Technologies for Advancing Big Data Science & Engineering

NSF - Methods to derive knowledge from data; infrastructure to manage, curate, and serve data; Approaches to education and workforce development

NIH - 1000 Genomes Project Data on Cloud

DOD - Data to Decisions - ~$250 million annually across military departments and ~25 million annually in the DARPA XDATA program

DOE - Establish the Scalable Data Management, Analysis and Visualization (SDAV) Institute within the SCIDAC program

USGS – Big Data for Earth System Science

This is just a piece of the puzzle - there are myriad other efforts throughout the Government!

Big Data Initiatives and Programs in NASA

Advanced Information Systems Technology (AIST) – Reduce the cost of evolving NASA information systems to support future Earth Observation missions and transform observations into information

Earth Science Data and Information System (ESDIS) – Process, archive and distribute Earth Science satellite data and data from airborne and field campaigns (via EOSDIS – the Earth Observing System Data and Information System)

Global Earth Observation System of Systems (GEOSS) – Collaborative, international effort to share and integrate Earth observation data. Joint effort with NOAA and the EPA.

Space Act Agreement – Collaborative effort with Cray, Inc. to explore development and application of low-latency big data systems

Planetary Data System (PDS) – Products from NASA planetary missions

Multi-Mission Archive at the Space Telescope Science Institute – Astronomical data archive and tools

Earth System Grid Federation – Public archive to support the IPCC Fifth Assessment Report. Collaboration with DOE.

What does Big Data have to do with you?

If you work in Earth Science,
you are in the world of Big Data

And it will only get bigger and more complex

NASA is trying to help!!

"Big Data"

What is it?

Why do you care?

- **NASA's Big Data**

What's out there?

Where is it?

How do you access/acquire it?

Major Categories of NASA Data

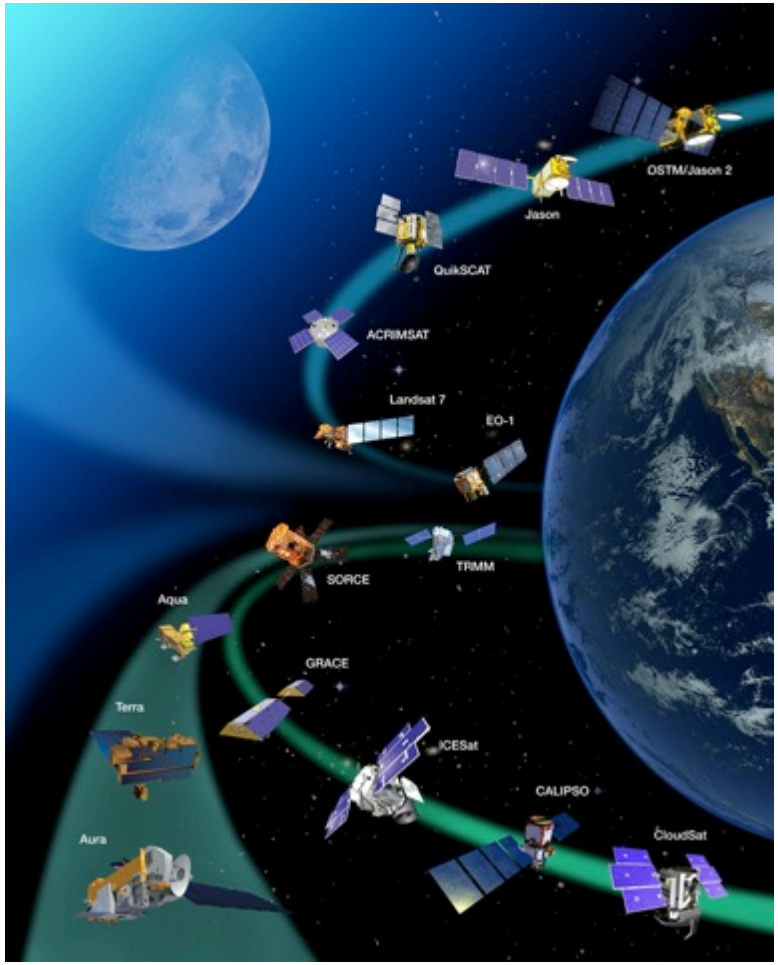
Satellite Remote Sensing Data

Assimilated Datasets (Validation Data)

Model Output

Climate Projections

Working with Satellite Remote Sensing Data from NASA



Finding it

Understanding it

Working with it

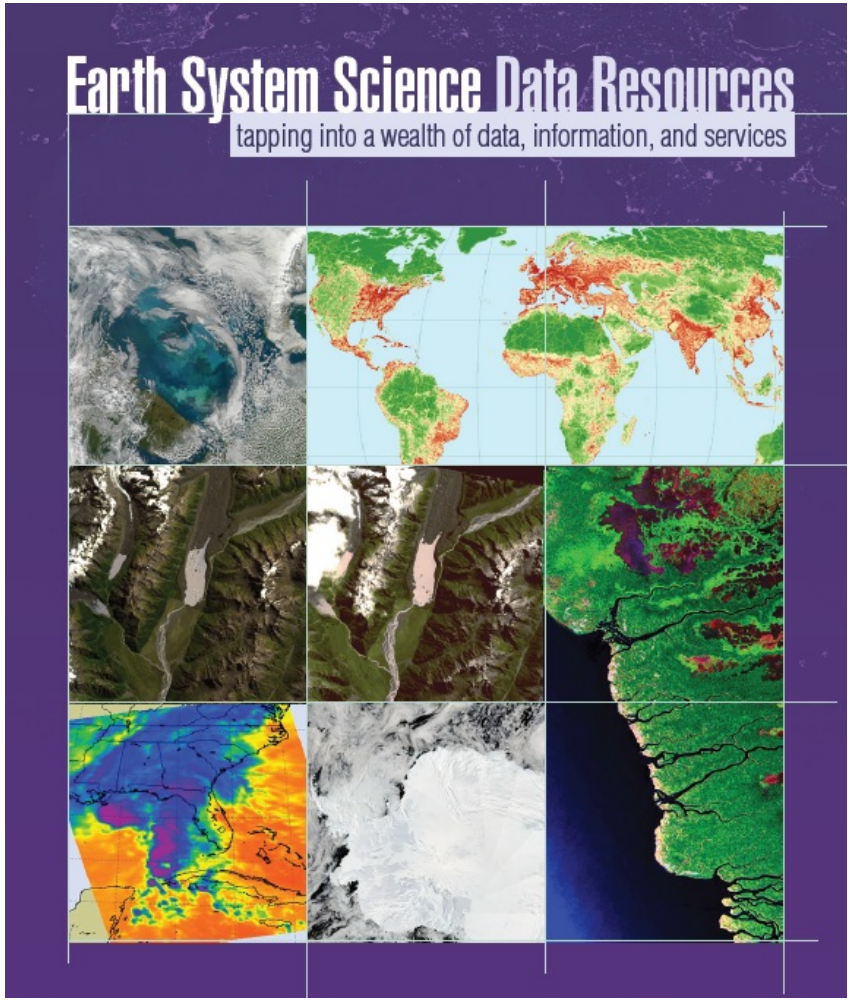
And most importantly…
Getting help!!!

Start Here!! Your guide to EOSDIS resources

Contains key information about…

- Sensors on NASA spacecraft
- Data and file formats
- NASA's 12 data centers and their data holdings
- Tools for data discovery and access
- Other sources for more detailed information

You will each receive a copy … and it is on the web!



http://earthdata.nasa.gov/library/earth-system-science-data-resources

NASA's Earth Science Data Centers

- Alaska Satellite Facility Synthetic Aperture Radar Data Center (ASF SDC)
- Crustal Dynamics Data Information System
- GSFC Earth Sciences Data and Information Services Center (GES DISC)
- Global Hydrology and Resource Center DAAC (GHRC DAAC)
- Langley Research Center Atmospheric Science Data Center (ASDC)
- Land Processes DAAC (LP DAAC)
- MODAPS Level 1 Atmosphere Archive and Distribution System (MODAPS LAADS)
- National Snow and Ice Data Center DAAC (NSIDC DAAC)
- Oak Ridge National Laboratory DAAC (ORNL DAAC)
- Ocean Biology Processing Group
- Physical Oceanography DAAC (PO DAAC)
- Socioeconomic Data and Applications Center (SEDAC)

The Tools available to you

The very top level…

- REVERB – A web-based client for discovering and ordering cross-discipline data from all twelve data centers' metadata holdings
- Global Change Master Directory (GCMD) – Allows users to discover and access more than 25,000 Earth science data sets and services
- Search and order within each data center

The Tools available to you

More detailed tools and services

- ~ 75 tools distributed across the centers for specialized access and manipulation of data sets

- Five categories
  - Search and Order
  - Data Handling
  - Subsetting & Filtering
  - Geolocation, Reprojection & Mapping
  - Data Visualization & Analysis

Finding data collections

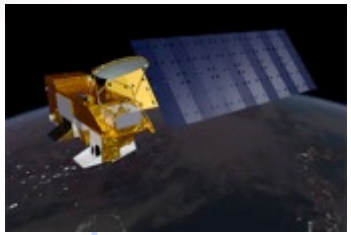
Key Factors

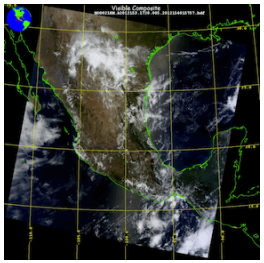
- What physical measurement do you want?
- What spatial coverage and resolution do you need?
- What temporal coverage and resolution do you need?

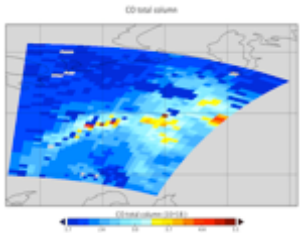
These will get you started, but you will often find that you need to understand more about how the data were collected and created

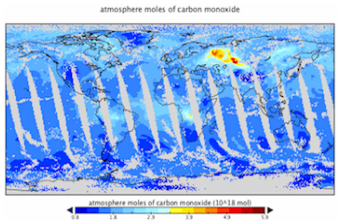
- Instrument characteristics
- Processing levels

Satellite data processing levels









Level 0 (L0)
raw data
packets

Level 1b (L1b)
calibrated
geolocated swaths

Level 2 (L2)
geophysical measurement
swaths

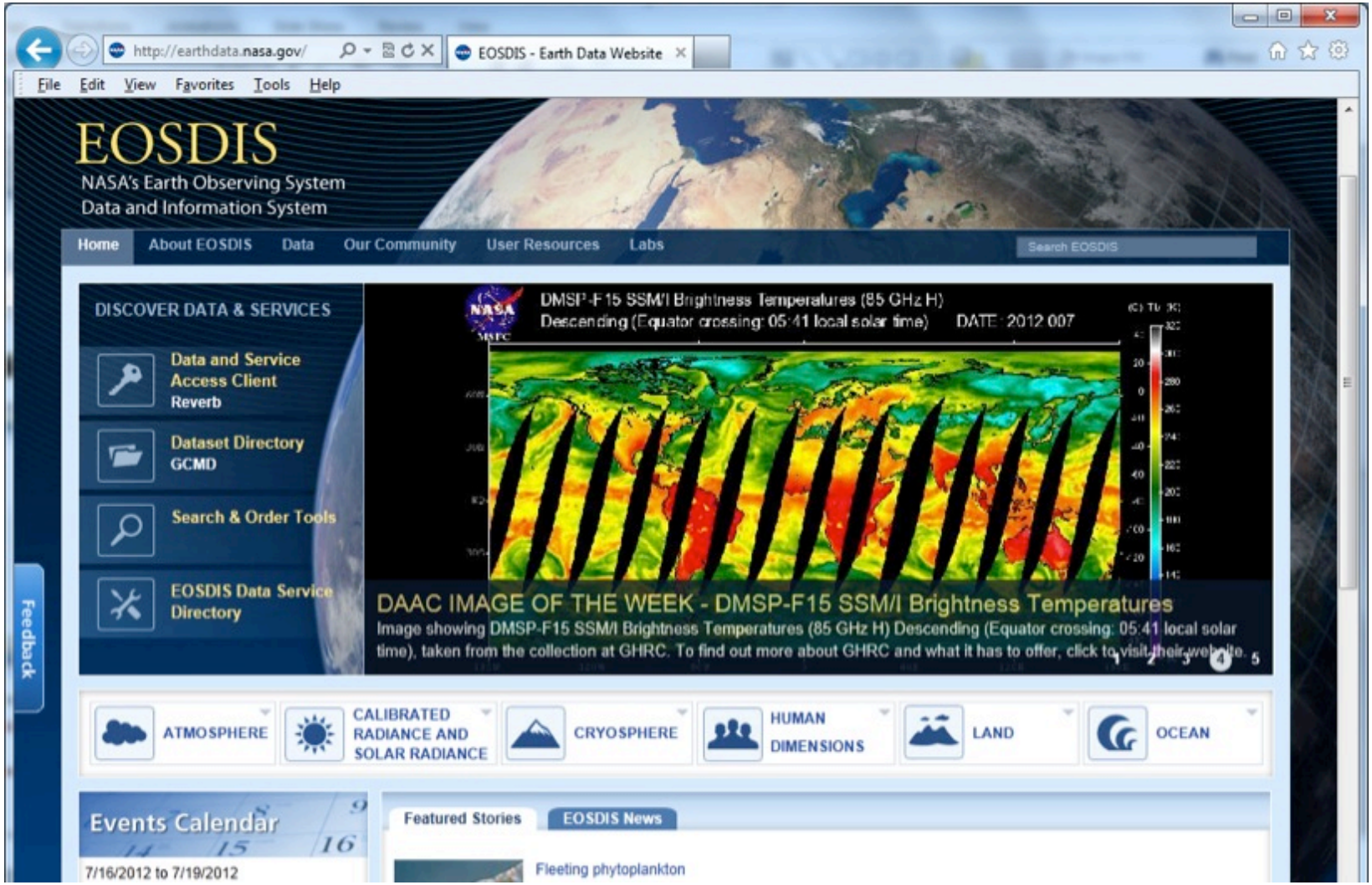
Level 3 (L3)
geophysical measurement
regular grids

Level 4 (L4)
model output
regular grids

EOSDIS Data Center

The EOSDIS Portal

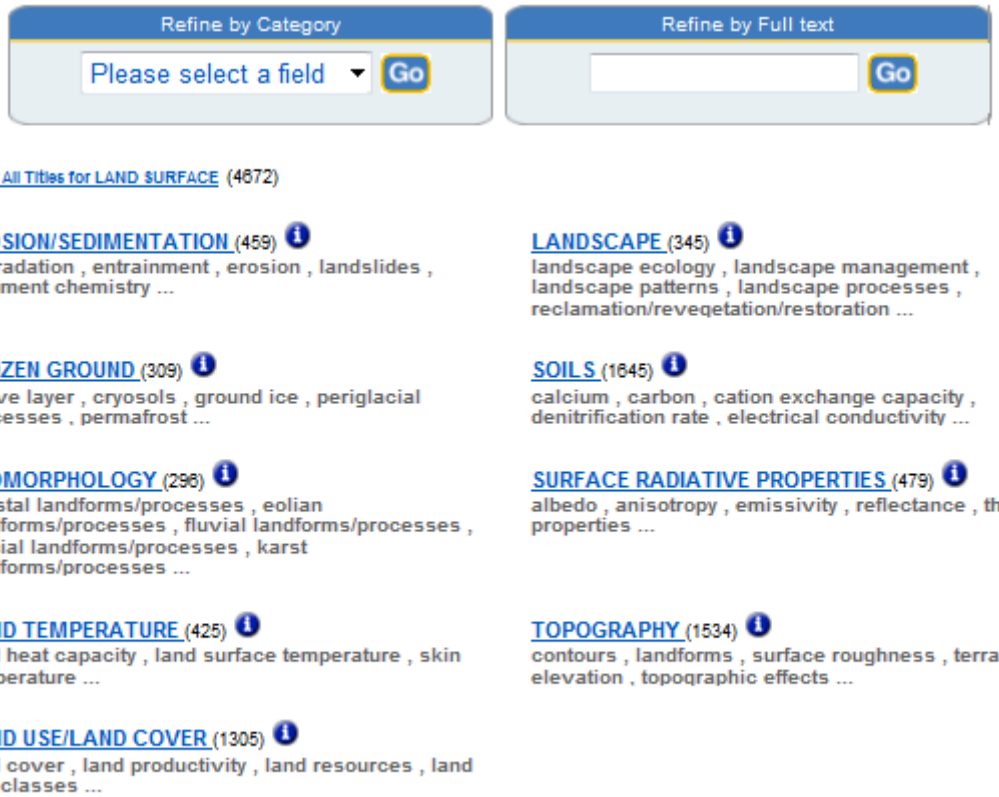
http://earthdata.nasa.gov



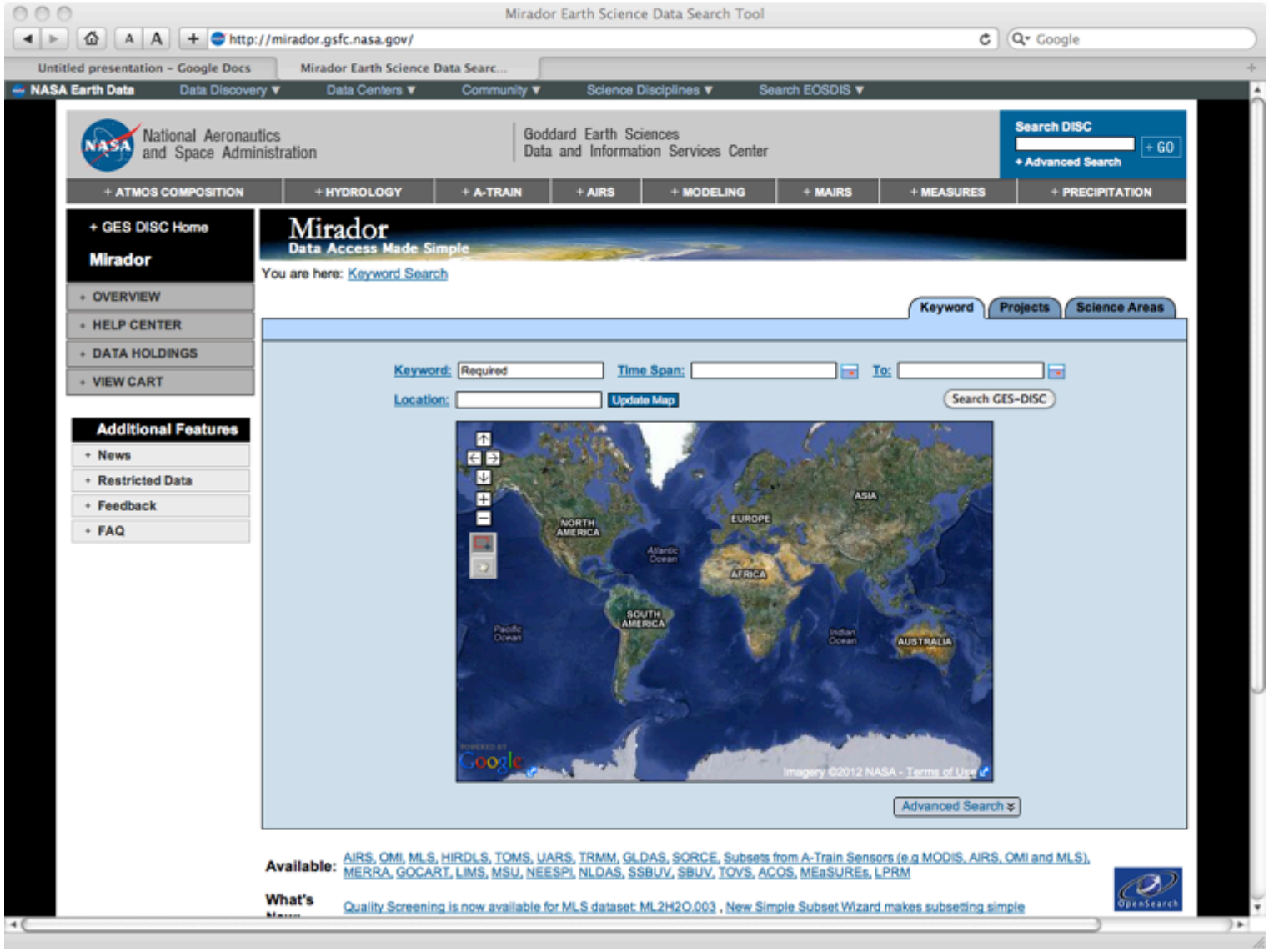
Reverb: Access All EOSDIS Data

Global Change Master Directory

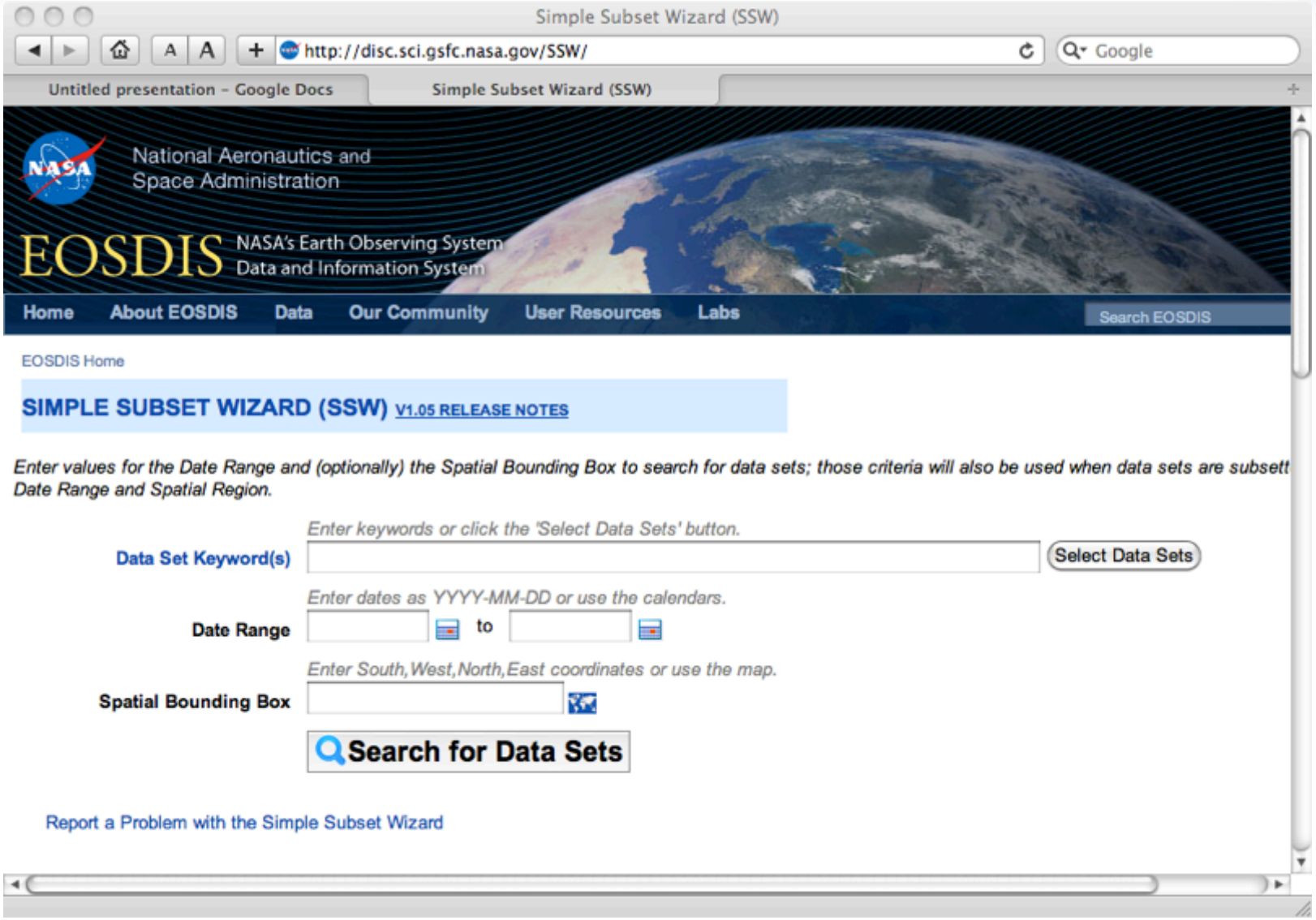
Drilling down into Land Surface data



Mirador: Search Goddard Earth Sciences Data and Information Services Center (GES DISC)



Simple Subset Wizard



Using Data

Data Formats

- HDF4, HDF5: advanced data formats requiring an API or tool
- HDF-EOS, HDF-EOS5: HDF with standard structures for geolocation
- netCDF: advanced data formats requiring an API or tool
- netCDF/CF-1: netCDF with standards for geolocation, time and vertical coordinates
- Some flat binary formats (ASCII and GRIB) and codes like BUFR

netCDF Tools with GUI

Panoply

- http://www.giss.nasa.gov/tools/panoply
- quick-view, shallow learning curve
- cross-platform
- reads some HDF (L2), lots of OPeNDAP

IDV (integrated data viewer)

- http://www.unidata.ucar.edu/software/idv/
- sibling product is McIDAS-V
- More flexible than Panoply
- A little harder to learn
- Reads many other formats, including OPeNDAP, but NOT HDF

non-GUI netCDF Tools

nco (netcdf command operators)

- Command-line software for doing math on netCDF files
- Useful utilities for manipulating (editing) netCDF files

ncl (netcdf command language)

- More geared toward plotting

Ferret: http://ferret.wrc.noaa.gov/Ferret/

- Geared toward plotting

Gridded Analysis and Display System (GrADS)

- http://www.iges.org/**grads**/
- Only works well with L3-L4 data
- Can read OPeNDAP
- Can read HDF with more effort

HDF Tools

HDFView

- http://www.hdfgroup.org/hdf-java-html/**hdfview/**

hdp (HDF4), h5dump (HDF5)

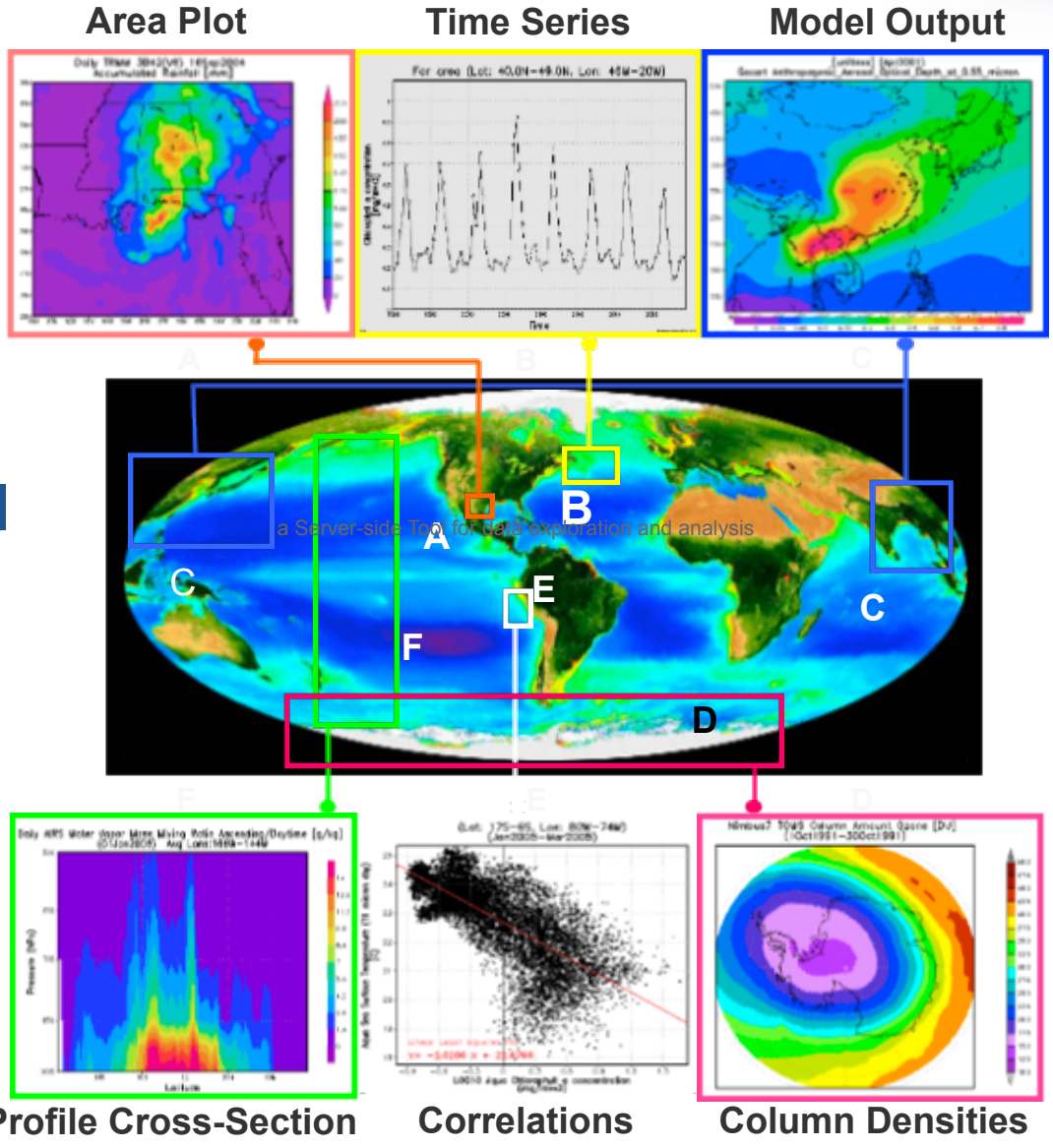
- Comes with HDF libraries
- Extracts data from HDF files into ASCII or flat binary

HDF data served through OPeNDAP “looks” like netCDF data to most clients

Visualizing and Analyzing Data

Giovanni

A Server-side Tool for data exploration and analysis



Giovanni Interface

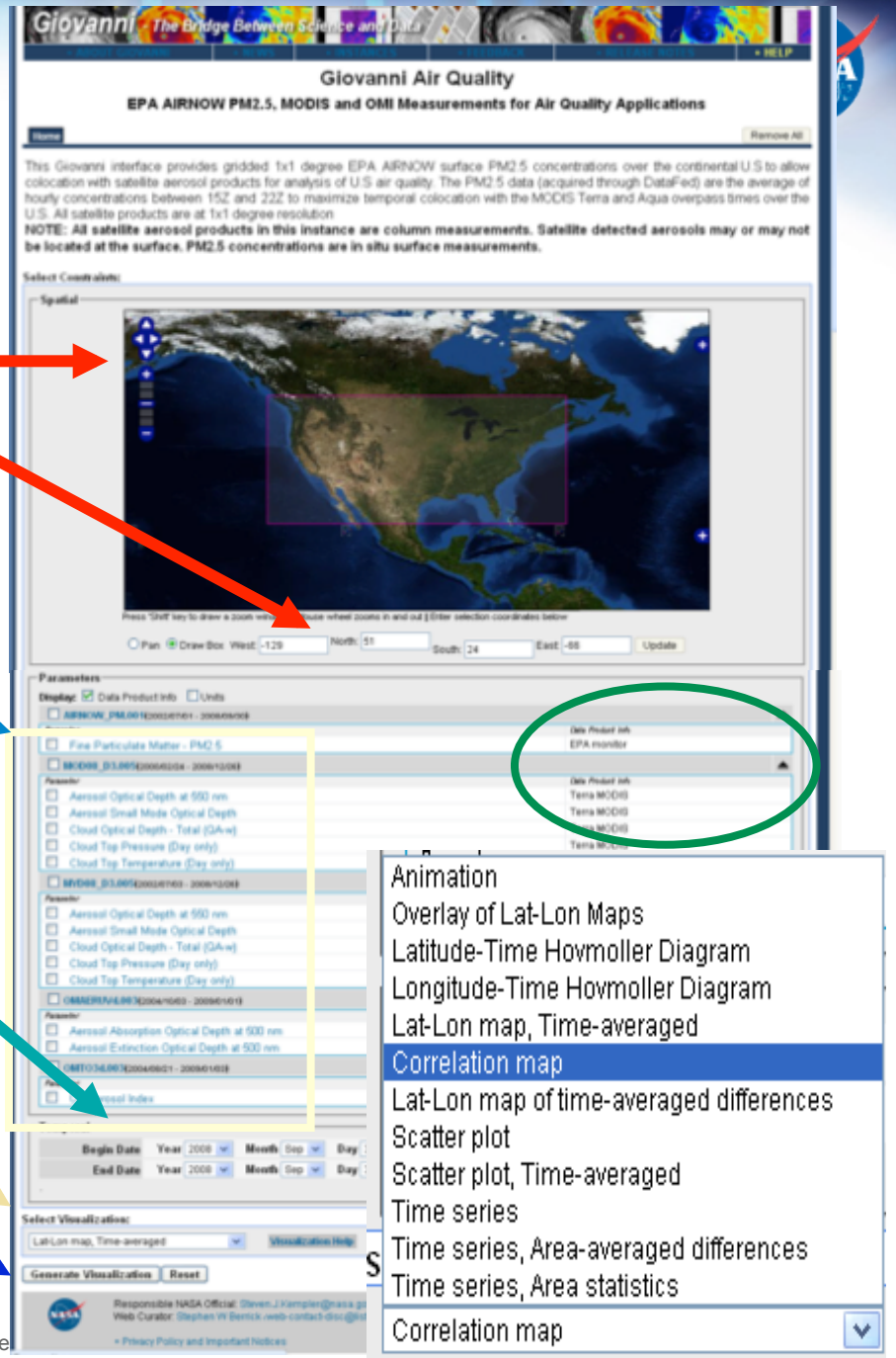
Select Area of interest

Select Variables

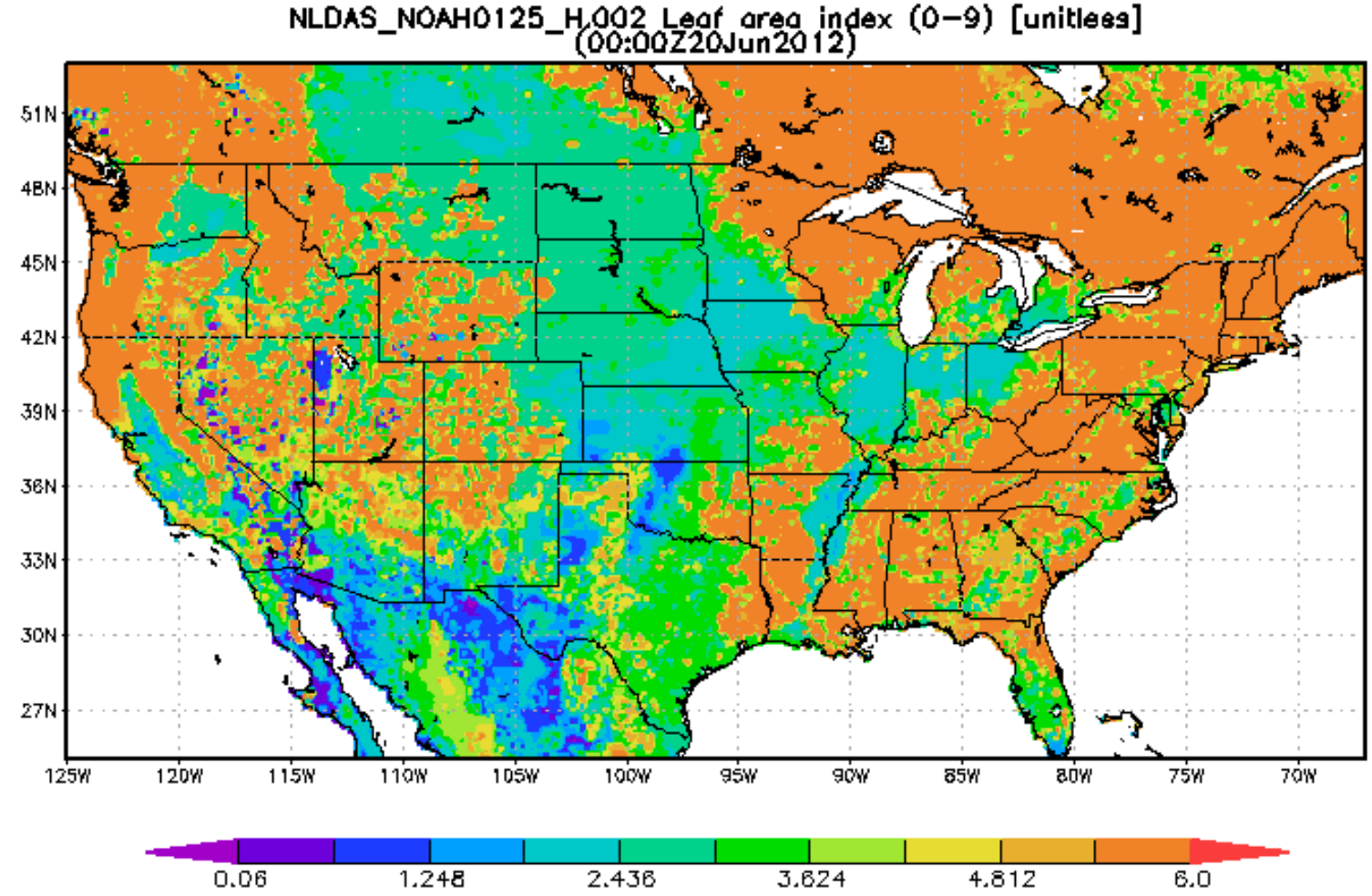
Select Time Period

Select Plot type

Generate Visualization



E.g., NLDAS Leaf Area Index



Giovanni Details

http://giovanni.gsfc.nasa.gov

- Mostly Level 3 data
- Mostly data from GES DISC

Several available functions

- Time averaging, spatial averaging, regression

Land Data in Giovanni

- Monsoon Asia Integrated Regional Study (MAIRS)
- Northern Eurasia Earth Science Partnership Initiative (NEESPI)
- North American Land Data Assimilation System (NLDAS)
- Global Land Data Assimilation System (GLDAS)

Model and Assimilation Data

- MERRA – Using 30 years of satellite data and the most advanced modeling and data assimilation techniques to create a reference data set
- ESGF – Consolidating access to the massive amount of model data being created for the next IPCC Assessment Report (AR5)

MERRA
Modern Era Retrospective Analysis for Research and Applications

What is MERRA?

NASA's ambitious effort to extract new information from 30 years of satellite observations using GEOS-5 Data Assimilation System

- New analysis technique for 4D representation of the state of the atmosphere
- Exploits NASA and NOAA rich satellite data holdings, additional satellite retrieval products, and conventional observations
- Provides data at higher spatial and temporal resolution

MERRA
Modern Era Retrospective Analysis for Research and Applications

Why is MERRA Important?

Provides invaluable data source for model assessment, climate & weather research, and hydrologic studies

- **Model Validation** – re-analysis datasets provide an improved representation of observational data to support assessment of model forecast skill
- **Scientific Research** – re-analysis datasets provide an improved derivation of geophysical parameters from observational data to support scientific study

MERRA
Modern Era Retrospective Analysis for Research and Applications

How do you get MERRA data?

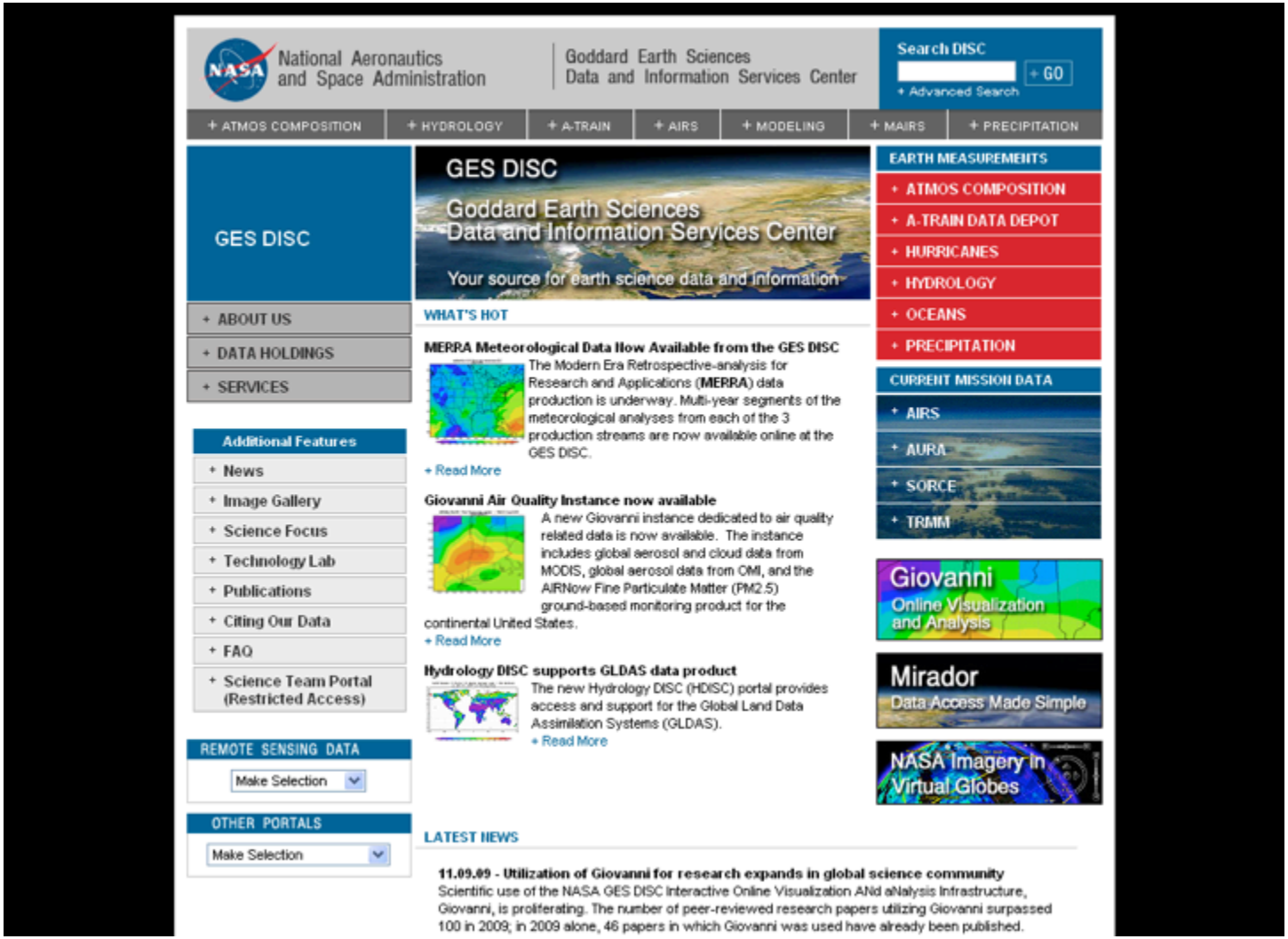
NASA/GSFC Global Modeling and Assimilation Office (GMAO)

- Understanding comparative merits of MERRA
  http://gmao.gsfc.nasa.gov/research/merra/prequel/view.php
- Finding and manipulating MERRA data with OpenDAP
  http://gmao.gsfc.nasa.gov/pubs/docs/da_Silva380.pdf

NASA Goddard Earth Sciences Data & Information Services Center (GES DISC)

- Accessing MERRA data via GES DISC
  http://daac.gsfc.nasa.gov/
- Re-analysis and observing system data manipulation with Giovanni
  http://giovanni.gsfc.nasa.gov/

Accessing MERRA via GES DISC



NASA Model and Observation Data on the Earth System Grid (ESG)

Earth System Grid Federation – a partnership of climate modeling centers

- Provide secure web-based distributed access to Coupled Model Intercomparison Project Phase 5 (CMIP5) model data that will be the basis for IPCC AR5
- Ability to search, download analyze and visualize this distributed data through one common web interface
- NCCS ESG node provides access to
  - Model data produced by the GMAO and GISS
  - CMIP5 data from NOAA and Center for Ocean-Land-Atmosphere Studies (COLA) produced with the National Centers for Environmental Prediction (NCEP) model
  - Selected MERRA datasets

NCCS ESG Node Home Page

And finally - NEX Data Holdings

These are datasets easily available to researchers with accounts on the NEX systems

Dataset	Variable	Volume	Location
NARR	Multiple	2.6TB	Pleiades, Lou
ASTER	DEM	256GB (gz)	Pleiades, Sandbox
CARBONTRACEKR	Multiple	701GB (gz)	Pleiades, Lou
CMIP	Multiple	2TB	Pleiades, Lou
CRU-NCEP	Multiple	1.3TB	Pleiades, Sandbox, Lou
GIMMS	Multiple	13GB	Pleiades, Lou
GTOPO30	Multiple	6GB	Pleiades, Sandbox, Lou
LANDSAT	Multiple	240TB (gz)	Pleiades, subset on Sandbox, Lou
MODIS	Multiple	30TB	Pleiades, subset on Sandbox, Lou
MSTMIP	Multiple	4TB	Pleiades, Sandbox, Lou
NACP	Multiple	500BG	Pleiades, Lou
NED	Multiple	26GB	Pleiades, Sandbox, Lou
PRISM	Multiple	1TB	Pleiades, Lou
SRTM	DEM	100GB	Pleiades, Sandbox, Lou
TOPS	Multiple	20TB	Lou
TOPS-CLIM	Multiple	10TB	Sandbox, Lou
MERRA	Multiple	59GB	Sandbox, Lou
NCDC	Multiple	17GB	Pleiades, Sandbox, Lou
NCEP	Multiple	21GB	Sandbox, Lou
TRMM	Multiple	1GB	Sandbox, Lou
OMI	Multiple	1GB	Sandbox, Lou
STATSGO	Multiple	1GB	Sandbox, Lou
WELD	Surface Reflectances	20TB	Pleiades
NADF3	Multiple	100GB	Pleiades, Lou
SIMS	Multiple	7TB	Sandbox, Lou

Thank You!!!

2012 Summer Short Course for Earth System Modeling and Supercomputing

Getting data files

- Search by time, space, and data collection attributes (satellite, instrument, processing level)
  - All of EOSDIS: Reverb, http://reverb.echo.nasa.gov
    - One stop shopping, but
    - Slower, more cumbersome for large data searches
  - Individual data centers
    - http://earthdata.nasa.gov/data/data-tools/search-and-order-tools
    - Optimized for the data center's data
  - Simple Subset Wizard: http://disc.gsfc.nasa.gov/SSW
    - Covers 10 of 12 datasets
    - Highly simplified user interface
    - *Hint*: try SSW first, then other search tools if you don't find what you need

Getting data files (con't)

Search tools will provide access to online URLs or ability to order data

Most data centers offer *dataset-specific* additional services

- Subsetting (spatial, by variable)
- MODIS Reprojection Tool
- Data Quality Screening Service (AIRS, MODIS L2)
- Reformat to netCDF
- Simple Subset Wizard offers some of these services
- OPeNDAP, Web Coverage Service, Web Map Service

Caveat #1: PAY ATTENTION TO DATA QUALITY!!

Most Level 1 and 2 data has quality control parameters

- Also, some L3, e.g., MODIS Land Surface Temperature

Failure to filter data according to quality flags can result in significant bias in your results

E.g., Data Quality Screening Service

- Only available for certain L2 atmospheric products at this point

Caveat #2
Daily Level 3 data is not a daily "average"!

The Diurnal Sampling Bias

Technically each L3 grid cell value is the average of L2 pixels falling in that cell for that day,

BUT...

In reality, it has contributions from only one or two satellite overpasses during the day*

- therefore about a 100-minute window during the day
- roughly the SAME 100-minute window every day for sun-synchronous satellites
- so averaging over weeks, months, years does not eliminate the Diurnal Sampling Bias